10-2016

Selecting Effective Examples to Train Students for Peer Review of Open-Ended Problem Solutions

Matthew Verleger Dr.
Embry-Riddle Aeronautical University, matthew.verleger@erau.edu

Kelsey J. Rodgers Dr.
Embry-Riddle Aeronautical University, rodgerk6@erau.edu

Heidi Diefes-Dux
Purdue University

Scholarly Commons Citation

Verleger, M., Rodgers, K. J., & Diefes-Dux, H. (2016). Selecting Effective Examples to Train Students for Peer Review of Open-Ended Problem Solutions. *Journal of Engineering Education, 105*(4). https://doi.org/10.1002/jee.20148

Running Heads:

- Selecting Effective Samples to Train Students for Artifact Peer Review
- Verleger, Rodgers, & Diefes-Dux

Selecting Effective Samples to Train Students for Artifact Peer Review

Matthew A. Verleger[a], Kelsey J. Rodgers[a], Heidi A. Diefes-Dux[b]
[a]Embry-Riddle Aeronautical University, [b]Purdue University

ABSTRACT

BACKGROUND

Students conducting peer review on authentic artifacts require training. In the training studied here, individual students reviewed (score and provide feedback on) a randomly selected prototypical solution to a problem. Afterwards, they are shown a side-by-side comparison of their review and an expert's review, along with prompts to reflect on the differences and similarities. Individuals were then assigned a peer team's solution to review.

PURPOSE

This paper explores how the characteristics of five different prototypical solutions used in training (and their associated expert evaluations) impacted students' abilities to score peer teams' solutions.

DESIGN/METHOD

An expert rater scored the prototypical solutions and 147 student teams' solutions that were peer reviewed using an eight item rubric. Differences between the scores assigned by the expert and a student to a prototypical solution and an actual team solution were used to compute a measure of the student's improvement as a peer reviewer from training to actual peer review. ANOVA

testing with Tukey's post-hoc analysis was done to identify statistical differences in improvement based on the prototypical solutions students saw during the training phase.

RESULTS

Statistically significant differences were found in the amount of error a student made during peer review between high and low quality prototypical solutions seen by students during training. Specifically, a lower quality training solution (and associated expert evaluation) resulted in more accurate scoring during peer review.

CONCLUSIONS

While students typically ask to see exemplars of "good solutions", this research suggests that there is likely greater value, for the purpose of preparing students to score peers' solutions, in students seeing a low-quality solution and its corresponding expert review.

Keywords: Peer review, Peer instruction, Model-eliciting activities

INTRODUCTION

Engineering courses are continually striving to include more authentic and open-ended problems as a means of providing students with engineering-like experiences to address the more difficult to achieve ABET criteria (Diefes-Dux, Moore, Follman, Zawojewski, & Imbrie, 2004; Judith S. Zawojewski, Hjalmarson, Bowman, & Lesh, 2008). Effective open-ended problem solving experiences require students to perform multiple iterations of revision, typically in response to constructive feedback (Verleger & Diefes-Dux, 2008). Providing written feedback, particularly on students' solutions (artifacts) to open-ended tasks, is time consuming. One way to decrease the burden on instructional staff, and to achieve other student learning benefits, is to engage students in peer review of each other's work. However, students need practice at peer review so that they understand the expectations for a high quality solution and can learn to provide

constructive feedback. The selection of samples for training students to conduct peer review should be intentional; however, instructor guidance in this regard is scarce.

At the large, mid-west, RU/VH institution where this study was conducted, mathematical modeling problems were used to engage students in authentic problem solving. In reviewing a large number of student solutions to these problems (Judith S. Zawojewski, Diefes-Dux, & Bowman, 2008), the research team recognized that students' first draft solutions were generally weak and that they would benefit from multiple iterations of feedback and revision (Verleger & Diefes-Dux, 2008). Achieving more than one iteration was a challenge when Teaching Assistants (TAs), at the time, were responsible for assessing 16 teams' solutions and each team's solution took upwards of 30 minutes to properly interpret and provide feedback (Cardella, Diefes-Dux, Oliver, & Verleger, 2009). In addition to this, TAs had other grading and classroom responsibilities.

Peer review was a means of achieving an additional iteration of feedback without increasing the TAs' workload. The students' peer review training was modeled after the current TA training to address concerns about the potential low quality of students' peer reviews (Verleger & Diefes-Dux, 2013). While TAs practiced assessing three to five sample student solutions, students only had the time to practice on one. Over time, this began to raise the question of what makes a good training sample. Samples were selected that, as a whole, were of lower quality so that students did not blatantly copy what they saw into their own solutions; though specific aspects were of higher quality to allow for some variability in what students saw during the review process. The question of impact of sample selection on students' ability to score and provide feedback during

peer review was not raised until researchers started examining student solutions to address research questions with regards to students' mathematical model development across iterations (Carnes, Diefes-Dux, & Cardella, 2011) and optimal matching of peer reviewers' skills to reviewees' needs for feedback (Verleger, Diefes-Dux, Ohland, Besterfield-Sacre, & Brophy, 2010). The question this study intends to address is "Does the quality of the sample solution used in training impact student error in subsequent scoring of another team's solution during peer review?". Answers to this question may help identify how students can be more effectively trained to accurately score peers' solutions.

LITERATURE REVIEW

Peer Review

Peer feedback and assessment are becoming more essential to engineering education, with some first-year engineering programs switching primarily to peer assessment because of the acknowledged benefits to students' professional skills development, potential for increasing students' motivation, and overall high desire for students taking responsibility (van Hattum-Janssen & Maria Lourenço, 2006). The use of peer review aligns with the recommendations and goals of IEEE (IEEE, 2007) and ASCE (American Society of Civil Engineers, 2008) to enhance engineers' peer feedback skills. With the increased student to teacher ratio, it is recognized that peer review is a useful tool for timely, formative feedback to students (O'Moore & Baldock, 2007; van Hattum-Janssen & Maria Lourenço, 2006).

Peer review can be a useful tool for helping students learn, but it is not without its challenges. Peers typically have a difficult time trusting fellow students and therefore using their peers' reviews (de Moreira, 2003; Eric Zhi-Feng Liu, Lin, Chi-Huang Chiu, & Shyan-Ming Yuan,

2001; K. J. Rodgers et al., 2015). This is an understandable concern, as research has shown that students typically assign a higher grade to their peers' solutions than a grader or expert would (K. J. Rodgers, Diefes-Dux, & Cardella, 2012). However, students have shown an ability to rank the quality of their peers' solutions similarly to experts, albeit with the assignment of higher rating values (Billington, 1997; Cheng & Warren, 1999). One option is to remove numeric scoring altogether from the peer review process and focus exclusively on written feedback (O'Moore & Baldock, 2007), though this approach effectively solves the problem by ignoring it. Even though teams may not rigorously use the peer feedback they receive, the review process has been shown to be effective in critically evaluating one's own solution (Ballantyne, Hughes, & Mylonas, 2002; Sitthiworachart & Joy, 2003) and developing the ability to critically evaluate others' work (Ballantyne et al., 2002; Boud, 2000; Guilford, 2001; Sitthiworachart & Joy, 2004).

Even through students struggle to effectively give and respond to peer reviews, multiple studies indicate that the quality of the products being submitted improved subsequent to the peer review process. Ballantyne et al. (2002) reported that the majority of their 939 first and second year survey respondents "agreed that peer assessment was an awareness-raising exercise because it made them consider their own work more closely, highlighted what they needed to know in the subject, helped them make a realistic assessment of their own abilities, and provided them with skills that would be valuable in the future." (p. 434) Similarly, Sitthiworachart and Joy (2003) indicated that 69% of first-year undergraduate students in computer science reported that they discovered mistakes in their own code while reviewing code written by their peers. Eighty percent of the students felt that seeing other students' work was helpful for their learning.

In addition to the immediate skills improvement provided by peer review, many researchers recognize the long-term benefits provided to reviewers. Giving effective constructive feedback is not only a struggle for first-year engineering students; it is also a challenge for undergraduate and graduate engineering teaching assistants (Cardella et al., 2009; Owens, 2011), engineers in industry (McCarthy & Garavan, 2001) as well as STEM professors/instructors (Carless, Salter, Yang, & Lam, 2011). Boud (2000) posited that the focus of assessment as a whole must be rethought to promote lifelong learning skills. Learning to perform peer review and to respond to formative feedback given via both peer and self-review are essential skills for success in a real work environment that does not assign an end-of-project grade. Teaching students how to accurately perform peer review and how to utilize constructive criticism for improvement is essential for their future careers.

Peer feedback skills are acknowledged to be important in STEM careers (Clough, 2004; Franklin, 2001; "IEEE - IEEE Code of Ethics," 2013) and an essential part of students' education (Accreditation Board for Engineering and Technology, 2013). These skills can only benefit from better training for students on giving accurate and high quality feedback and on interpreting and responding to such feedback.

Peer Review Training Methodologies

Peer assessment and feedback training strategies within engineering, business, and writing tend to focus on qualitative, formative feedback over numeric scoring (O'Moore & Baldock, 2007), train students to give feedback within a well-structured model (Harms & Roebuck, 2010; O'Moore & Baldock, 2007), and involve continuous peer feedback (Lam, 2010; O'Moore &

Baldock, 2007). Alternative modes for providing feedback are possible. Some modes enable more formative feedback, such as using an online tool or other technology-based environment for feedback outside of the classroom (Lam, 2010).

One peer feedback training program, used in writing courses, found success with a model that consists of four steps: (1) clarifying intentions, (2) identifying problem areas, (3) explaining problem areas, and (4) giving direction to improve shortcomings (Davis & Foster, 2002). This four step process was revised from a model proposed by Min (2005). The training consists of a three-week in-class training program in which instructors describe this model of feedback, explain the importance of feedback, provide methods for receiving feedback, and ask students to apply this model to give feedback. Davis and Foster acknowledged that the implementation of feedback training through an online environment enables students to give more formative feedback because there are fewer time constraints than in-class. However, the training process still consumes a significant amount of time.

Another common tool used in conducting peer review is called “Calibrated Peer Review” (CPR), the result of an NSF funded project built by researchers at UCLA (Calibrated Peer Review, 2015; Robinson, 2001). The tool is purposefully designed to train students for peer review of essays using expertly selected or developed samples. Prior to participating in peer review, students must first evaluate those expert samples. While the process and tool have demonstrated success at engaging students in peer review (Prichard, 2005; Reynolds, 2008) or at improving the quality of submitted solutions (Gunersel, Simspon, Aufderheide, & Wang, 2008; Hartberg, Gunersel, Simspon, & Balester, 2008), little attention has been paid to the selection or development process

for those training samples. The design of the CPR system utilizes three calibration stages with high, medium, and low quality samples. Students review all three as part of their calibration process. Because all three are included, there is no mechanism for differentiating which samples actually induce learning.

Most studies of peer review training focus on qualitative feedback, eschewing the quantitative aspects of peer review. As students do not typically perceive the quantitative feedback provided by peers as accurate (and are partially correct in that assertion), this study focused on trying to improve the accuracy of the quantitative markings as a means of beginning to address students' overall negative perception regarding peer review. It is hoped that, by improving the accuracy of the quantitative markings, students will become more receptive to the qualitative feedback they receive, judging the entire package of feedback they receive as more accurate.

Model-Eliciting Activities

The mathematical modeling problems used in this study were Model-Eliciting Activities (MEAs). As a pedagogical tool, MEAs have been explored within the engineering context (Diefes-Dux, Bowman, Zawojewski, & Hjalmarson, 2006; Verleger et al., 2010) and were found to be effective mechanisms for helping students to develop a deeper understanding of engineering content. MEAs have primarily been studied in the context of very large first-year introductory engineering courses (Diefes-Dux & Cardella, 2008; Diefes-Dux & Imbrie, 2008; K. J. Rodgers et al., 2015; Verleger et al., 2010; Verleger & Diefes-Dux, 2013), though analysis of their effectiveness has also been done in higher level discipline-specific contexts (Bowman & Siegmund, 2008; Self et al., 2008; Yildirim, Shuman, & Besterfield-Sacre, 2010). A significant

portion of the research on their use has centered around developing formative and summative feedback mechanisms that are reliable across a variety of evaluators, including peer reviewers (Diefes-Dux, Zawojewski, & Hjalmarson, 2010; Diefes-Dux, Zawojewski, Hjalmarson, & Cardella, 2012).

Within the MEA context, the benefits of including both quantitative and qualitative feedback in the peer review process are clear. For the reviewee, the use of a single, consistent research-derived grading rubric (Diefes-Dux et al., 2010) across model development and feedback iterations highlights for students increases and decreases in the quality of their team's solutions and the changes that need to be made. The quantitative scoring also provides a gauge for whether the written feedback is referring to big or small issues in the solution. The primary benefit for the reviewer is that, through the use of the rubric, the reviewer can come to better appreciate the features of high quality solutions. The methodology of this paper is centered on using a training process similar to that of the CPR system to identify the kinds of training samples that are most effective at reducing the amount of quantitative error students make during peer review. The end goal is to quickly and more effectively train students for peer review, thereby reducing the error they make and the perception that peer feedback is less valuable than instructor feedback.

METHODS

What follows is a description of the context in which this study took place, followed by a closer examination of the impact different training samples had on how accurately students performed peer review.

Course

The data for this study was collected during the fall 2008 semester of a required introductory problem-solving and engineering computer tools course at a large mid-west, RU/VH university. The course enrollment was limited to students in the first-year engineering program.

Participants

The lead author has extensive MEA related experience. He evaluated the solutions of 147 teams from a purposefully sampled cross-section of the course designed to isolate instructor differences and course meeting times (Verleger, 2009); the net effect was a statistically random sample from across a large class. This study is an extension of that work, using portions of the same data set but looking at other attributes of the student work. The 147 teams consisted of 584 students. After removing those students who did not participate in both the training and peer review phases, the study population was reduced to 449 students. As can be seen in

Table 1, the gender and ethnic demographics of the study population (N=449) is representative of the students in the whole class (N=1164).

Table 1. Gender and Ethnicity of Students

	Whole Class		Study Population	
N	1164		449	
Ethnicity / Sex	Female	Male	Female	Male
Caucasian American	16.2%	60.3%	17.1%	60.1%
Asian American	1.9%	6.9%	1.8%	6.9%
Spanish American[a]	0.5%	2.2%	0.7%	2.0%
Other American	1.0%	1.7%	0.9%	1.6%
African American	0.9%	1.7%	0.9%	1.3%
American Indian	0.3%	0.5%	0.2%	0.0%
Unspecified or International	1.7%	4.1%	1.6%	4.9%
	22.5%	77.5%	23.2%	76.8%
	100%		100%	

[a] Ethnicity category used by university in 2008

Model Eliciting Activity

Selected Problem In this study, student team responses to the Purdue Paper Plane Challenge (PPPC) MEA were analyzed. The PPPC MEA, a variant of which is described by Wood et al. (2008), requires that students develop a procedure to assist the judges of a paper airplane contest in ranking the award winning teams for four categories: Most Accurate, Best Floater, Best Boomerang, and Best Overall, based on given measurements of time in air, distance from target, and length of throw for multiple throws on a straight path and a boomerang path. Student teams responses were in the form of 1 to 3 page memo detailing the team's procedure for solving the problem.

Implementation Highlights The administration of the MEA followed the sequence shown in Figure 1. The details of this sequence are described in Verleger et al. (Verleger et al., 2010). During the implementation of the PPPC MEA, students' attention was drawn explicitly to the rubric used to assess their work twice. First, prior to the start of the PPPC MEA, expectations for developing a high quality, generalizable, and share-able solution were set in lecture, with particular emphasis on understanding how the MEA Rubric dimensions (discussed below) relate to developing a high quality solution. Second, following TA feedback on the teams' first drafts, time was spent in lecture helping students understand how to interpret feedback in terms of its relationship to the MEA Rubric items. So prior to peer review training and peer review, the students had seen and heard about the MEA Rubric dimensions on three occasions, in the two aforementioned lecture instances and while reviewing the TAs' feedback on first team draft.

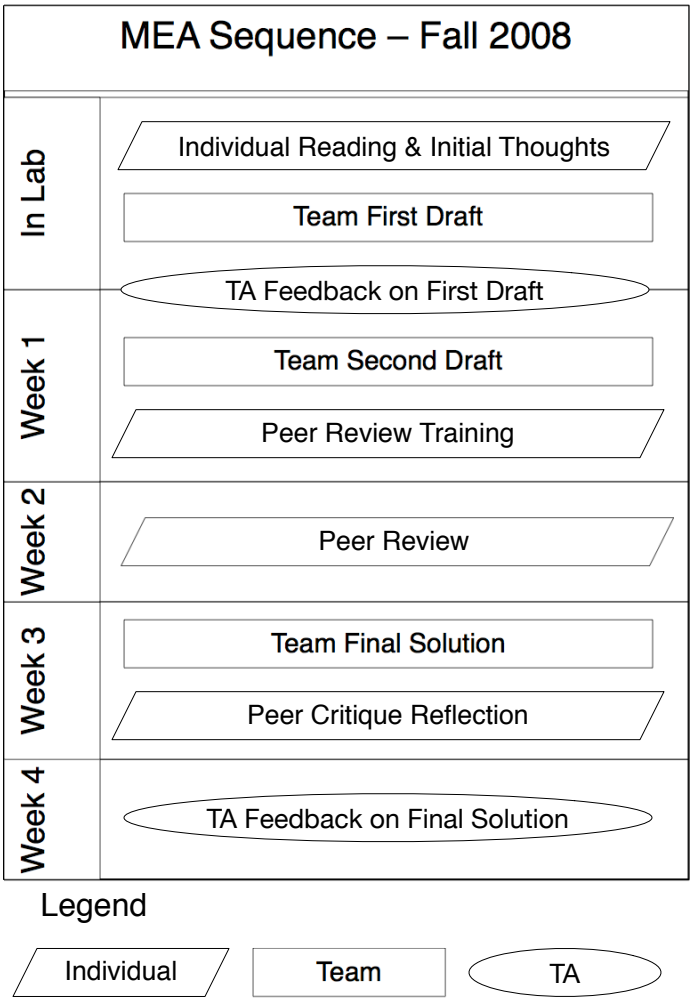

Figure 1. Fall 2008 MEA Sequence

Peer Review Training and Peer Review Prior to peer review, students were trained in the use of the MEA Rubric (described below); the students evaluated a randomly selected sample from a pool of five prototypical student team solutions. Instructors and researchers were blind to the results of this training assignment until after the MEA was complete. Samples were selected from a database of student solutions from prior semesters. The solutions were explicitly chosen

to be of generally low to moderate quality in order to mimic the solutions students would likely see while conducting their actual peer review. After being selected, the samples were updated to remove any identifying information and to reflect any changes to the MEA, including updating results to accommodate changes in the MEA itself. Care was taken to ensure that the essence of the solutions did not change and that the samples could reasonably come from a team of peers currently enrolled in the class. After students submitted their evaluation of the samples, they were shown their review next to the third author's review of that same sample. The students were asked to reflect on how they could improve their ability to evaluate and provide feedback on an MEA solution. Following this training, students participated in a double-blind peer review of an actual solution developed by a team of their peers. Students used the same MEA Rubric in complete this review.

MEA Rubric A full discussion on the development, reliability, and validity of the MEA Rubric can be found in Diefes-Dux, Zawojewski, and Hjalmarson. (2010). The rubric, for this semester, was used to assess student work along three dimensions: Mathematical Model, Re-Usability & Modifiability, and Audience (Share-ability). Each dimension contained numeric and free response feedback items. The numeric components are shown in Table 2. The free response prompts asked explicitly for items such as a summary of the mathematics used or recommendations for improving the rationales. These prompts were intended to help students more deeply engage in the review process by directing their attention to specific aspects of a team's solution so that better feedback would be generated.

Each of the eight quantitative MEA Rubric items consisted of either a true/false prompt or a set of mutually exclusive descriptive prompts. Point values were assigned to each possible selection. The prompts and point values can be seen in Table 2, grouped by rubric dimension. The score for each dimension was calculated as the minimum of the items in that dimension; the overall score was calculated as the minimum score of the three dimension scores. In classroom practice, the taking of minimums to is done to encourage continuous broad-spectrum improvement. This is also a philosophical stance by the instructors – a team's solution is only as good as the weakest dimension.

As an example, assume a reviewer selected "False" for No Progress (4 points), "A procedure somewhat addresses the complexity of the problem or contains embedded errors." for Mathematical Model Complexity (2 points), "False" for Data Usage (3 points), and "True" for Rationales (4 points). The Mathematical Model Dimension score is then the minimum of 4 (No Progress), 2 (Mathematical Model Complexity), 3 (Data Usage), and 4 (Rationales), resulting in a score of 2 for the Mathematical Model Dimension. The overall score is then calculated as a minimum of the three dimensional scores. For the example, the best possible overall score is a 2. The score is anchored by the Mathematical Model Dimension score of 2 as no other dimension provides a lower possible score.

Table 2. MEA Rubric – Numerical Items

Dim.	Item Label	Full Item Wording	Points	
Mathematical Model	No Progress	No progress has been made in developing a model. Nothing has been produced that even resembles a poor mathematical model. For example, simply rewriting the question or writing a "chatty" letter to the client does not constitute turning in a product.	True	0
			False	4
	Mathematical Model Complexity	The procedure fully addresses the complexity of the problem.	4	
		A procedure moderately addresses the complexity of the problem or contains embedded errors.	3	
		A procedure somewhat addresses the complexity of the problem or contains embedded errors.	2	
		Does not achieve the above level.	1	
	Data Usage	The procedure takes into account all types of data provided to generate results OR justifies not using some of the data types provided.	True	4
			False	3
	Rationales	The procedure is supported with rationales for critical steps in the procedure.	True	4
			False	3
Re-Usability/Modifiability	Re-Usability/ Modifiability	The procedure not only works for the data provided but is clearly re-usable and modifiable. Re-usability and modifiability are made clear by well articulated steps and clearly discussed assumptions about the situation and the types of data to which the procedure can be applied.	4	
		The procedure works for the data provided and might be re-usable and modifiable, but it is unclear whether the procedure is re-usable and modifiable because assumptions about the situation and/or the types of data that the procedure can be applied to are not clear or not provided.	3	
		Does not achieve the above level.	2	
Audience (Share-ability)	Results	Results from applying the procedure to the data provided are presented in the form requested.	True	4
			False	1
	Audience Readability	The procedure is easy for the client to understand and replicate. All steps in the procedure are clearly and completely articulated.	4	
		The procedure is relatively easy for the client to understand and replicate. One or more of the following are needed to improve the procedure: (1) two or more steps must be written more clearly and/or (2) additional description, example calculations using the data provided, or intermediate results from the data provided are needed to clarify the steps.	3	
		Does not achieve the above level.	2	
	Extraneous Information	There is no extraneous information in the response.	True	4
			False	3

Evaluation of Training Samples and Team Solutions

For this study, twelve different sets of evaluations, listed chronologically in **Error! Reference source not found.**, were conducted by four different evaluator groups. The third author developed the training samples and the "Comparison to Expert" evaluations used during the training phase of the peer review (Evaluation E1). The TAs and students then participated in the MEA as part of their normal course work (Evaluations E2-E5). After the MEA was complete, the first author evaluated the training samples as a means of establishing inter-rater reliability (analysis method described below) (Evaluation E6). He then evaluated all three drafts of the 147 teams in the sample (Evaluations E7-E9). Finally, to determine intra-rater reliability, 42 teams were randomly selected to be re-evaluated: 14 teams' first drafts (Evaluation E10), another 14 teams' second drafts (Evaluation E11), and another 14 teams' final solutions (Evaluation E12) (analysis method described below).

Table 3. Chronological Timeline of Evaluations

Evaluation	Work Being Evaluated	Evaluator
E1	Training Samples	Third author
E2	First Draft Solution	TAs
E3	Training Samples	Students
E4	Second Draft Solution	Students
E5	Final Solution	TAs
E6	Training Samples	First Author
E7	First Draft Solution	First Author
E8	Second Draft Solution	First Author
E9	Final Solution	First Author
E10	14 Random First Draft Solutions	First Author
E11	14 Random Second Draft Solutions	First Author
E12	14 Random Final Solutions	First Author

Inter-Rater Reliability of Evaluation

Because the students' training was primarily built around the sample evaluations of the third author (Evaluation E1), the first author had to establish inter-rater reliability with the third author's evaluations. To identify the inter-rater reliability, the five training samples were evaluated by the first author (Evaluation E6) and scores were compared to those established by the third author for use in the training. Of the 60 total markings (5 samples * [8 items + 3 dimensions + an overall score]), both authors were in perfect agreement on 52 items (87%) and within one level on 7 items (12%). Only one Audience Readability item (2%) had a difference of two levels.

Intra-Rater Reliability of Evaluation

Evaluation of all three drafts of the solutions by all 147 teams (Evaluations E7-E9) took the first author approximately 10 weeks. Upon completion, he felt it necessary to establish intra-rater reliability as a means of validating that his evaluations were sufficiently consistent throughout the 10-week span. Forty-two (42) teams were selected and their drafts were evaluated a second time by the first author (Evaluations E10-E12). Spearman's Rho correlation coefficients (α = 0.05) were calculated between the author's original evaluation (Evaluations E7-E9) and the second evaluation (Evaluations E10-E12), as seen in Table 4. While there is room for improvement, all 12 items were considered strong enough for the first author's evaluation to be considered acceptably reliable. This aligns with the interpretations described by Cohen (1988) and Corder and Foreman (2009) stating that, for the behavioral sciences, correlation values greater than 0.5 are considered strong, while values greater than 0.3 should be considered moderate.

Table 4. Intra-Rater Reliability Results

Dimension/Item/Overall Score	Intra-rater Reliability Correlations N = 42 Spearman Rho Correlation	Sig. (2-tailed *p*-value)
Mathematical Model Dimension	0.66	0.000
No progress has been made in developing a model.	1.00	0.000
The procedure fully addresses the complexity of the problem.	0.63	0.000
The procedure takes into account all types of data provided to generate results OR justifies not using some of the data types provided.	0.78	0.000
The procedure is supported with rationales for critical steps in the procedure.	0.69	0.000
Re-Usability / Modifiability Dimension	0.97	0.000
The procedure not only works for the data provided but is clearly re-usable and modifiable.	0.97	0.000
Audience (Share-ability) Dimension	0.80	0.000
Results from applying the procedure to the data provided are presented in the form requested.	0.77	0.000
The procedure is easy for the client to understand and replicate.	0.75	0.000
There is no extraneous information in the response.	0.61	0.000
Overall Score	0.77	0.000

Analysis

A Training Error (TE) was calculated as the sum of the differences between a student's training score and the expert's corresponding score on each of the eight rubric items. Likewise, a Peer Review Error (PRE) was calculated as the sum of the differences between the student's peer review score and the expert's corresponding score on each of the eight rubric items. An improvement score (IScore) was calculated as TE – PRE. A positive TE – PRE value is desired as this means the student made less error during the peer review phase than the training phase.

RESULTS & DISCUSSION

Results

Means and standard deviations of the IScores, as well as F statistics and significance values for each training sample (A to E) are given in **Error! Reference source not found.**. ANOVA testing with Tukey's post-hoc analysis (α=0.05) (King, Rosopa, & Minium, 2011) was conducted to detect the impact each of the five training samples had on the IScore. Results of the ANOVA are given in **Error! Reference source not found.**, with Tukey's analysis being shown in Table 6.

In Table 6, each row represents a statistically homogenous subset (p<=0.05). For example, for the Audience Readability item, training samples A, B, and C are statistically significantly similar to each other in their effect on the IScore. Students who received samples A, B, or C during their training had an average *decrease* of 0.421 in the amount of error they then made during peer review. Read another way, their average IScore went up by 0.421. Samples D and E are also statistically significantly similar to each other, but students who received samples D and E during training averaged had an *increase* of 0.442 in peer review error; their average IScore went down by 0.442. Their quantitative scores were less accurate during peer review than during their training. Finally, Samples A, B, and C are statistically significantly different from Samples D and E ($p < 0.05$). Samples A, B, and C each received an expert's Audience Readability score of 2 out of 4 while Samples D and E had scores of 3 out of 4.

Table 5. Mean, Standard Deviation, ANOVA F and Significance Values of Student IScores

Sample ID		**A**	**B**	**C**	**D**	**E**	F	Sig.
n		91	89	81	89	99	df=4,444	
No Progress	μ	–0.13	–0.27	0.2	0.09	–0.08	1.263	0.284
	σ	1.26	1.79	1.38	1.47	1.61		
Mathematical Model Complexity	μ	0.62	–0.33	0.46	0.26	–0.49	24.924	0.000

	σ	1.19	1.08	1.27	1.01	1.19		
Data Usage	μ	–0.05	–0.06	–0.09	–0.12	–0.01	0.34	0.851
	σ	0.80	0.79	0.84	0.74	0.81		
Rationales	μ	0.4	0.12	0.31	0.16	0.24	2.313	0.057
	σ	0.71	0.75	0.79	0.73	0.73		
Re-Usability/ Modifiability	μ	0.27	0.34	0.28	0.25	0.17	0.545	0.703
	σ	1.02	0.93	0.90	1.04	1.01		
Results	μ	–0.59	1.31	1.44	0.24	0.58	17.088	0.000
	σ	1.88	2.00	1.83	2.19	1.80		
Audience Readability	μ	0.4	0.48	0.38	–0.4	–0.48	40.283	0.000
	σ	0.80	0.80	0.88	0.77	0.85		
Extraneous Information	μ	–0.35	–0.34	–0.3	–0.26	–0.28	0.391	0.815
	σ	0.65	0.69	0.74	0.72	0.68		

Table 6. Training Sample and Change in Reviewer Error Homogenous Subsets (*n*=449)

	Expert's Scores for Sample Work					Average IScore (Decrease in Reviewer Error)[a]						
Sample ID	A	B	C	D	E	A	B	C	D	E	Group	Possible Range
n	91	89	81	89	99	91	89	81	89	99		
No Progress	4	4	4	4	4	–0.13	–0.27	0.20	0.09	–0.08	–0.044	+4 to –4
Mathematical Model Complexity	1		1	1		0.62		0.46	0.26		0.448	+3 to –3
		2			2		–0.33			–0.49	–0.414	
Data Usage	3	4	3	3	3	–0.05	–0.06	–0.09	–0.12	–0.01	–0.064	+1 to –1
Rationales	3	3	3	3	3	0.40	0.12	0.31	0.16	0.24	0.245	+1 to –1
Re-Usability/ Modifiability	2	2	2	2	2	0.27	0.34	0.28	0.25	0.17	0.260	+2 to –2
Results		1	1				1.31	1.44			1.372	+3 to –3
		1			1		1.31			0.58	0.926	
				1	1				0.24	0.58	0.419	
	4					–0.59					–0.590	
Audience Readability	2	2	2			0.40	0.48	0.38			0.421	+2 to –2
				3	3				–0.40	–0.48	–0.442	
Extraneous Information	4	4	4	4	4	–0.35	–0.34	–0.30	–0.26	–0.28	–0.306	+1 to –1

[a] Negative numbers represent an *increase* in error from the Training to the Peer Review stage

[b] Rows represent homogenous subsets (i.e., for the Results item, Sample A is statistically significantly different than Samples B, C, D, & E, while Samples B and C, B and D, and D and E are not statistically significantly different from each other, each at the $p <= 0.05$ level).

Discussion

What follows is a discussion of the rubric item results found in Table 6 that revealed interesting characteristics about the training samples.

Mathematical Model Complexity Item

As shown in Table 6, students who saw a higher quality mathematical model (and the associated expert feedback) during their training showed an increase in their average peer review error. Students who saw a lower quality sample (and the associated expert feedback) had a statistically significant decrease in their average peer review error.

The training samples provided similar mathematical models, but differed in that the higher quality samples tended to include explicit, but general (i.e., non-mathematical), definitions of the four paper airplane awards that were independent of their mathematical procedural steps. For example, Sample E states, “The competition for Best Floater should be defined as the paper plane that travels the shortest distance in the longest amount of time.” This sample then goes on to describe how to numerically determine the best floating plane based on that definition. In contrast, Sample D states, “To determine the Best Floater, the average time for each team will be taken from the three attempts and the highest one overall will be awarded "Best Floater."” This represented Sample D’s entire procedural approach for that award, with no independent definition of what makes a plane a good floater. This difference is subtle and is likely not definitive enough to result in the difference in peer review error.

Five other possible sources of difference among the samples were investigated but yielded no significant findings. These are summarized as:

- The assigned training sample was similar to the students' own Draft 1 – numerical analysis of the first author's scores found no statistically significant correlation. This assumes quantitative scores reflect the nature of the solution.
- The assigned training sample was similar to the students' own Draft 2 – numerical analysis of the first author's scores found no statistically significant correlation. This assumes quantitative scores reflect the nature of the solution.
- The assigned training sample was similar to the actual solution assigned for peer review – numerical analysis of the first author's scores found no statistically significant correlation.
- The text provided by the expert induced the change – content analysis found no clear indicators that the text was drastically different among all samples and in fact it was often the same exact text.
- There was a slight but not statistically significant difference in character count of the expert's Mathematical Model feedback. Samples A, C, and D (Expert Rating 1) had an average character count of 3272 versus Samples B and E with an average character count of 2879. The character count of each sample is shown in Table 7.

Table 7. Mathematical Model Expert Feedback Character Count

Sample	A	B	C	D	E
Character Count	3430	3079	3165	3222	2679

It is likely that small combinations of all of these items (and possibly others) are the root cause of the difference in peer review error.

Data Usage Item

As noted in the Data Usage section of Table 6, Sample B received a higher score compared to the other four samples but yielded no significant impact on peer review error, with error increasing regardless of the provided sample. Further investigation revealed that Sample B's level 4 was based on their use of all the prescribed data, though the expert noted in the feedback, "While your team has used all of the different data types for Best Boomerang and Best Overall, their use is not necessarily meaningful." Further, the expert's feedback about data usage on all of the samples was limited and tended to focus more on clarifying if the data being used was from the straight or boomerang throws. In the solutions being peer reviewed, 52.6% of the solutions had no clear indications of what data was being used (e.g., straight throw data versus boomerang throw data), 27.5% of the solutions omitted an entire data type (e.g., not using length of throw from either the straight throws or boomerang throws), and an additional 31.3% omitted at least one column of data from their analysis (e.g., not using length of throw from the straight throw but using the length of throw from the boomerang throw), none of which were issues that the expert feedback on the training samples discussed. The mismatch between what the expert feedback focused on in the training samples and what students actually had problems with in their solutions is the most likely cause for the slight increase in peer review error.

Results Item

The design of an MEA intentionally affords students' ability to self-assess their procedure (Diefes-Dux, Hjalmarson, Miller, & Lesh, 2008); students are required to include the results of applying their procedure to the provided data. The nature of the PPPC MEA required that students identify both the final paper airplane awardees as well as provide the quantitative values leading to the assignment of those awards. Sample A included all of the requested information, while Samples B – E did not. As a consequence, Sample A's expert feedback with regards to the results item only included a minor discussion about significant figures, while Samples B – E all included more explicit requests for quantitative results as well as identification of the winning teams. As evidenced by the increase in error for students assigned Sample A versus a decrease in error for students assigned Samples B-E, the explicit request for quantitative results in the feedback provided by the expert played a key role in reducing peer review error.

Audience Readability Item

The purpose of the readability item is to measure how well constructed the solution is and how easily the results (evaluated in the item above) can be reproduced. The idea is to determine how easy it is for the reader (as proxy for the client) to use the solution, keeping in mind that the reader had no involvement in the development of the procedure.

Samples A, B, and C had lower readability than Samples D and E and resulted in decreased error, while Samples D and E had higher readability than Samples A, B, and C and resulted in an increase in error. Much like the results item, there is a clear split indicating that students trained using samples that are of lower readability tended to be more aware of what makes a low quality solution, while individuals trained on samples of higher readability were less able to identify when a solution is considered low quality.

Rationales Item and Re-usability/Modifiability Item

The rationales and re-usability/modifiability items both saw decreases in reviewer error across all samples. Because there was no variability across the samples used in training, it is difficult to make comparisons. It should be noted that the score given for both rubric items was the lowest possible score for each of the respective items. As such, the expert feedback was targeted at describing general attributes associated with each item, as opposed to providing specific feedback for helping someone improve the components already present in their work.

Extraneous Information Item

The extraneous information item is designed to ensure that, while students should be explaining their thought process and the steps to their solution, they should also be focused on presenting their solution in a brief and efficient document. Common attributes that are considered extraneous are discussions of how to use various computer tools (e.g., MATLAB, Excel), how to calculate common statistical measures (e.g., mean, standard deviation), and restatements of large sections of the original problem text. None of the five training samples included extraneous information. The lack of extraneous information in the samples meant the expert did not discuss extraneous information in the feedback and the samples all received full marks. As a consequence, error during peer review went up, indicating that students did not sufficiently understand what constituted extraneous information.

CONCLUSIONS AND TRAINING SAMPLE SELECTION AND DEVELOPMENT

Peer feedback and assessment are becoming more essential to engineering education, with some first-year engineering programs trying to switch primarily to peer assessment because of the acknowledged student benefits (van Hattum-Janssen & Maria Lourenço, 2006). Having students

participate in training using samples similar to the solutions they will be reviewing provides them a low-stakes opportunity to practice developing a critique before undertaking peer review. In the face of larger classes and increasing student-teacher ratios, this approach can provide a low-resource means of training students to participate in peer review and a valuable way for continuing to give students timely formative feedback (O'Moore & Baldock, 2007; van Hattum-Janssen & Maria Lourenço, 2006).

Based on the authors' prior experiences, they recommend that instructors, whenever possible, use prior students' work as a starting place for samples as opposed to having faculty or TAs generate samples. While faculty can generate their own solutions, extra care must be taken to prevent generated solutions from being "too good", as faculty and teaching assistants often naturally use more sophisticated language than students are likely to encounter.

By using previous student solutions as a starting place, realistic solutions can be developed for use in peer review training, though care must be taken to effectively de-identify these samples to comply with privacy regulations. Using prior student work allows instructors the opportunity to potentially select from a wide variety of solutions that may be better able to meet specific objectives, such as exposing students to a unique approach to thinking about the problem. When selecting samples, instructors are encouraged to select samples that are representative of the types of work being submitted and that expose common mistakes that can be addressed through clear expert feedback.

Regardless of instructors' other goals, one likely goal is to reduce the amount of error students make during the peer review process. While some researchers suggest avoiding the error issue altogether by removing the numeric evaluation component of peer review (O'Moore & Baldock, 2007), we suggest a more critical training regimen focused on improving students' skills rather than avoiding students' potential shortcomings.

As this research demonstrates, selecting lower quality solutions *and* providing appropriately detailed expert feedback on that lower quality solution for students to compare against was found to reduce the amount of error students made during peer review as compared to the amount of error made in training. This is exemplified by the Mathematical Model Complexity, Results, and Readability rubric items shown in Table 6. Despite the tendency for students to request to see exemplars of superior solutions, this research suggests that seeing lower quality samples and the corresponding feedback about why that sample is considered lower quality may be more beneficial to their learning. The detail provided through the expert feedback becomes a clear indicator of what is expected for each rubric item, providing helpful guidance as students move into peer review.

The authors recommend the following process for instructors when selecting Samples:

1. Examine a wide variety of student work.
2. Identify common problems that need to be addressed.
3. Select samples that demonstrate these common problems. As needed, combine pieces of multiple solutions to generate a single sample, being careful to watch for continuity issues and language differences across the pieces.

4. Write expert feedback that highlights the problems and the corrective action needed to fix them. Faculty are reminded that this is exemplary feedback, not notes to the students undergoing training about why a sample was graded the way it was.

LIMITATIONS AND FUTURE WORK

This study provides insight into improving students' evaluations during peer review, though there are some limitations. First, this study is contextualized within the study of Model-Eliciting Activities. It would strengthen these conclusions to explore the described training model in other contexts and with more variable samples and other evaluation rubrics.

Another limitation is that the training samples lacked sufficient diversity in rubric scores. The expert's evaluations of each of the rubric items were all binary ratings, with only two different scores being assigned to any given item. The study would be strengthened by utilizing a more diverse set of training samples that might better elucidate the relationship between sample quality and reviewer error. It is an open question whether high-quality samples with appropriately detailed feedback, including praise on what common mistakes were avoided and what elements should be retained for the next draft, could produce similar improvements to low-quality samples.

An exploration of the alignment between the written feedback and numeric scores is also necessary. Numerous studies have explored the value of written feedback, while others have explored the numeric evaluations, but few studies have explored the relationship between those two aspects of evaluation. Some initial research with a mixed-methods approach was conducted

by Rodgers, et al. (2012), but this study also states much more mixed-methods research is needed.

Finally, one of the goals of this work is to improve the accuracy of peer quantitative markings as a means of bolstering the perceived validity of peer evaluations by students. These results generate two new research questions that now must be explored. First, is the reduction in error provided by using low quality training samples with detailed expert feedback enough to produce sufficiently accurate peer reviews? If so, how do teams respond to that feedback and what can be done to improve their response to higher quality feedback?

ACKNOWLEDGEMENTS

This work was made possible by a grant from the National Science Foundation (EEC 0835873 and EEC 1264005). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Accreditation Board for Engineering and Technology. (2013). *Criteria for Accrediting Programs in Engineering*. Baltimore, MA: ABET, Inc.

American Society of Civil Engineers. (2008). *2007-2008 Policies & Priorities*. Washington, D.C.: ASCE Washington Office. Retrieved from http://www.asce.org/files/pdf/pressroom/2007_2008PoliciesPriorities.pdf

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing Procedures for Implementing Peer Assessment in Large Classes Using an Action Research Process. *Assessment & Evaluation in Higher Education*, *27*(5), 427–441. http://doi.org/10.1080/0260293022000009302

Billington, H. L. (1997). Poster presentations and peer assessment: novel forms of evaluation and assessment. *Journal of Biological Education*, *31*(3), 218–220. http://doi.org/10.1080/00219266.1997.9655566

Boud, D. (2000). Sustainable Assessment: Rethinking assessment for the learning society. *Studies in Continuing Education*, *22*(2), 151–167. http://doi.org/10.1080/713695728

Bowman, K. J., & Siegmund, T. (2008). Designing Modeling Activities for Upper-Level Engineering Classes. In J. S. Zawojewski, H. Diefes-Dux, & K. Bowman (Eds.), *Models and Modeling in Engineering Education: Designing Experiences for All Students* (pp. 93–110). Rotterdam, The Netherlands: Sense Publishers.

Calibrated Peer Review. (2015). Calibrated Peer Review. Retrieved February 28, 2015, from http://cpr.molsci.ucla.edu

Cardella, M. E., Diefes-Dux, H. A., Oliver, A., & Verleger, M. A. (2009). Insights into the Process of Providing Feedback to Students on Open-Ended Problems. In *American Society for Engineering Education 2009 National Conference & Exposition* (pp. 14.742.1–14.742.18). Austin, TX. Retrieved from https://peer.asee.org/insights-into-the-process-of-providing-feedback-to-students-on-open-ended-problems

Carless, D., Salter, D., Yang, M., & Lam, J. (2011). Developing sustainable feedback practices. *Studies in Higher Education*, *36*(4), 395–407. http://doi.org/10.1080/03075071003642449

Carnes, M. T., Diefes-Dux, H. A., & Cardella, M. E. (2011). Evaluating Student Responses In Open-Ended Problems Involving Iterative Solution Development In Model-Eliciting Activities. In *American Society for Engineering Education 2011 National Conference & Exposition*. Vancouver, B.C., Canada.

Cheng, W., & Warren, M. (1999). Peer and Teacher Assessment of the Oral and Written Tasks of a Group Project. *Assessment & Evaluation in Higher Education*, *24*(3), 301–314. http://doi.org/10.1080/0260293990240304

Clough, G. W. (2004). The engineer of 2020: Visions of engineering in the new century. *National Academy of Engineering, Washington*.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Corder, G. W., & Foreman, D. I. (2009). *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*. Hoboken, N.J.: John Wiley & Sons Inc.

Davis, B. D., & Foster, P. (2002). Performance Documentation . *Business Communication Quarterly* , *65* (2), 108–114. http://doi.org/10.1177/108056990206500212

de Moreira, D. A. (2003). No Title. *Education and Information Technologies*, *8*(1), 47–54. http://doi.org/10.1023/A:1023926308385

Diefes-Dux, H. A., Bowman, K. J., Zawojewski, J. S., & Hjalmarson, M. A. (2006). Quantifying Aluminum Crystal Size Part 1: The Model-Eliciting Activity. *Journal of STEM Education*, *7*(1 & 2), 51–63.

Diefes-Dux, H. A., & Cardella, M. E. (2008). Formative Feedback: Impacting the Quality of First-Year Engineering Student Work on Modeling Activities. National Science Foundation (NSF) - EEC 0835873.

Diefes-Dux, H. A., Hjalmarson, M. A., Miller, T. K., & Lesh, R. A. (2008). Model-Eliciting Activities for Engineering Education. In J. S. Zawojewski, H. Diefes-Dux, & K. Bowman (Eds.), *Models and Modeling in Engineering Education: Designing Experiences for All Students* (pp. 17–36). Rotterdam, The Netherlands: Sense Publishers.

Diefes-Dux, H. A., & Imbrie, P. K. (2008). Modeling Activities in a First-Year Engineering Course. In J. S. Zawojewski, H. Diefes-Dux, & K. Bowman (Eds.), *Models and Modeling in Engineering Education: Designing Experiences for All Students* (pp. 55–92). Rotterdam, The Netherlands: Sense Publishers.

Diefes-Dux, H. A., Moore, T., Follman, D., Zawojewski, J. S., & Imbrie, P. K. (2004). Interactive Session - Model-Eliciting Activities: A Framework For Posing Open-Ended Engineering Problems. *34th ASEE/IEEE Frontiers in Education Conference*. Savannah, GA.

Diefes-Dux, H. A., Zawojewski, J. S., & Hjalmarson, M. A. (2010). Using Educational Research in the Design of Evaluation Tools for Open-Ended Problems. *International Journal of Engineering Education*, *26*(4), 807–819.

Diefes-Dux, H. A., Zawojewski, J. S., Hjalmarson, M. A., & Cardella, M. E. (2012). A Framework for Analyzing Feedback in a Formative Assessment System for Mathematical Modeling Problems. *Journal of Engineering Education*, *101*(2), 375–406. http://doi.org/10.1002/j.2168-9830.2012.tb00054.x

Eric Zhi-Feng Liu, Lin, S. S. J., Chi-Huang Chiu, & Shyan-Ming Yuan. (2001). Web-based peer review: the learner as both adapter and reviewer. *IEEE Transactions on Education*, *44*(3), 246–251. http://doi.org/10.1109/13.940995

Franklin, J. (2001). The Importance of Peer Reviewing. *The Scientific World JOURNAL*, *1*, 23–24. http://doi.org/10.1100/tsw.2000.18

Guilford, W. H. (2001). Teaching Peer Review and the Process of Scientific Writing. *Advances in Physiology Education*, *25*(3), 167–175.

Gunersel, A. B., Simspon, N. J., Aufderheide, K. J., & Wang, L. (2008). Effectiveness of Calibrated Peer Review for improving writing and critical thinking skills in biology

undergraduate students. *Journal of the Scholarship of Teaching and Learning*, *8*, 25–37.

Harms, P. L., & Roebuck, D. B. (2010). Teaching the Art and Craft of Giving and Receiving Feedback . *Business Communication Quarterly* , *73* (4), 413–431. http://doi.org/10.1177/1080569910385565

Hartberg, Y., Gunersel, A. B., Simspon, N. J., & Balester, V. (2008). Development of Student Writing in Biochemistry Using Calibrated Peer Review. *Journal of the Scholarship of Teaching and Learning*, *2*, 29–44.

IEEE. (2007). IEEE Envisioned Future. Retrieved from http://www.ieee.org/portal/cms_docs_iportals/iportals/aboutus/envisioned_future.pdf

IEEE - IEEE Code of Ethics. (2013). Retrieved August 6, 2013, from http://www.ieee.org/about/corporate/governance/p7-8.html

King, B. M., Rosopa, P. J., & Minium, E. W. (2011). *Statistical Reasoning in the Behavioral Sciences* (6th ed.). John Wiley & Sons, Inc.

Lam, R. (2010). A peer review training workshop: Coaching students to give and evaluate peer feedback. *TESL Canada Journal*, 114–127. Retrieved from http://www.teslcanadajournal.ca/index.php/tesl/article/view/1052

McCarthy, A. M., & Garavan, T. N. (2001). 360° feedback process: performance, improvement and employee career development. *Journal of European Industrial Training*, *25*(1), 5–32. http://doi.org/10.1108/03090590110380614

Min, H.-T. (2005). Training students to become successful peer reviewers. *System*, *33*(2), 293–308. Retrieved from http://www.sciencedirect.com/science/article/pii/S0346251X05000187

O'Moore, L. M., & Baldock, T. E. (2007). Peer Assessment Learning Sessions (PALS): an innovative feedback technique for large engineering classes. *European Journal of Engineering Education*, *32*(1), 43–55. http://doi.org/10.1080/03043790601055576

Owens, G. (2011). Transforming undergraduate structural engineering education in the 21st Century. *The Structural Engineer*, *89*(2).

Prichard, J. R. (2005). Writing to learn: an evaluation of the calibrated peer review™ program in two neuroscience courses. *Journal of Undergraduate Neuroscience Education : JUNE : A Publication of FUN, Faculty for Undergraduate Neuroscience*, *4*, A34–9. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3592621&tool=pmcentrez&rendertype=abstract

Reynolds, J. (2008). Calibrated Peer Review Assignments in Science Courses: *Journal of College Science Teaching*, 60–66.

Robinson, R. (2001). Calibrated Peer Review™. *The American Biology Teacher*. http://doi.org/10.1662/0002-7685(2001)063[0474:CPR]2.0.CO;2

Rodgers, K., Diefes-Dux, H. A., Cardella, M. E., & Fry, A. S. (2012). First-year engineering students' peer feedback on open-ended mathematical modeling problems. In H. A. Diefes-Dux, M. E. Cardella, & A. Fry (Eds.), *2012 Frontiers in Education Conference Proceedings* (Vol. 0, pp. 1–6). Seattle, WA: IEEE. http://doi.org/10.1109/FIE.2012.6462467

Rodgers, K. J., Diefes-Dux, H. A., & Cardella, M. E. (2012). AC 2012-3820 : The Nature of Peer Feedback from First-Year Engineering Students on Open-Ended Mathematical Modeling Problems. San Antonio, TX: American Society for Engineering Education.

Rodgers, K. J., Horvath, A. K., Jung, H., Fry, A. S., Diefes-Dux, H. A., & Cardella, M. E. (2015). Students' Perceptions of and Responses to Teaching Assistant and Peer Feedback. *Interdisciplinary Journal of Problem-Based Learning*, *9*(2). http://doi.org/10.7771/1541-5015.1479

Self, B. P., Miller, R. L., Kean, A., Moore, T. J., Ogletree, T., & Schreiber, F. (2008). Important student misconceptions in mechanics and thermal science: Identification using Model-Eliciting Activities. In *Proceedings - Frontiers in Education Conference, FIE*. http://doi.org/10.1109/FIE.2008.4720595

Sitthiworachart, J., & Joy, M. (2003). Deepening Computer Programming Skills by Using Web-based Peer Assessment. In *4th Annual Conference of the LTSN Centre for Information and Computer Sciences*. NUI Galway, Ireland.

Sitthiworachart, J., & Joy, M. (2004). Effective Peer Assessment for Learning Computer Programming. *9th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education* . Leeds, United Kingdom.

van Hattum-Janssen, N., & Maria Lourenço, J. (2006). Explicitness of criteria in peer assessment processes for first-year engineering students. *European Journal of Engineering Education*, *31*(6), 683–691. http://doi.org/10.1080/03043790600911779

Verleger, M. A. (2009). *Analysis of an Informed Peer Review Matching Algorithm and its Impact on Student Work on Model-Eliciting Activities*. *Engineering Education*. Purdue University, West Lafayette, IN.

Verleger, M. A., & Diefes-Dux, H. A. (2008). Impact of Feedback and Revision on Student Team Solutions to Model-Eliciting Activities. In *American Society for Engineering Education 2008 National Conference & Exposition*. Pittsburgh, PA.

Verleger, M. A., & Diefes-Dux, H. A. (2013). A Teaching Assistant Training Protocol for Improving Feedback on Open-Ended Engineering Problems in Large Classes. In *2013 ASEE Annual Conference* (pp. 23.121.1–23.121.12). Atlanta, GA. Retrieved from https://peer.asee.org/a-teaching-assistant-training-protocol-for-improving-feedback-on-open-ended-engineering-problems-in-large-classes

Verleger, M. A., Diefes-Dux, H. A., Ohland, M. W. M. W., Besterfield-Sacre, M., & Brophy, S. (2010). Challenges to Informed Peer Review Matching Algorithms. *Journal of Engineering Education*, *99*(4), 397–408. http://doi.org/10.1002/j.2168-9830.2010.tb01070.x

Wood, T., Hjalmarson, M. A., & Williams, G. (2008). Learning to Design in Small Group Mathematical Modeling. In J. S. Zawojewski, H. Diefes-Dux, & K. Bowman (Eds.), *Models and Modeling in Engineering Education: Designing Experiences for All Students* (pp. 187–212). Rotterdam, The Netherlands: Sense Publishers.

Yildirim, T. P., Shuman, L., & Besterfield-Sacre, M. (2010). Model-Eliciting Activities : Assessing Engineering Student Problem Solving and Skill Integration Processes. *International Journal of Engineering Education*, *26*, 831–845.

Zawojewski, J. S., Diefes-Dux, H. A., & Bowman, K. J. (2008). Models and Modeling in Engineering Education: Designing Experiences for All Students . Sense Publishers.

Zawojewski, J. S., Hjalmarson, M. A., Bowman, K. J., & Lesh, R. A. (2008). A Modeling Perspective On Learning And Teaching In Engineering Education. In J. S. Zawojewski, H. Diefes-Dux, & K. Bowman (Eds.), *Models and Modeling in Engineering Education: Designing Experiences for All Students* (1st ed., pp. 1–16). Rotterdam, The Netherlands: Sense Publishers.

Authors

Matthew A. Verleger is an Associate Professor of Engineering Fundamentals at Embry-Riddle Aeronautical University, 600 S. Clyde Morris Blvd., Daytona Beach, FL 32128; mverleger@gmail.com

Kelsey J. Rodgers is an Assistant Professor of Engineering Fundamentals at Embry-Riddle Aeronautical University, 600 S. Clyde Morris Blvd., Daytona Beach, FL 32128; rodgersk@erau.edu

Heidi A. Diefes-Dux is a Professor of Engineering Education at Purdue University, 701 W. Stadium Ave., West Lafayette, IN 47907; hdiefes@purdue.edu